Is Quality Control Pointless?

Markus Krause
Telefonica Alpha
Barcelona, Spain
markus@icsi.berkeley.edu

Farhad M. Afzali
University of Nebraska at
Omaha, NE, United States
mafzali@unomaha.edu

Simon Caton
National College of
Ireland, Dublin, Ireland
simon.caton@ncirl.ie

Margeret Hall
University of Nebraska at
Omaha, NE, United States
mahall@unomaha.edu

Abstract

Intrinsic to the transition towards, and necessary for the success of digital platforms as a service (at scale) is the notion of human computation. Going beyond 'the wisdom of the crowd', human computation is the engine that powers platforms and services that are now ubiquitous like Duolingo and Wikipedia. In spite of increasing research and population interest, several issues remain open and in debate on largescale human computation projects. Quality control is first among these discussions. We conducted an experiment with three different tasks of varying complexity and five different methods to distinguish and protect against constantly underperforming contributors. We illustrate that minimal quality control is enough to repel constantly underperforming contributors and that this is constant across tasks of varying complexity.

1. Introduction

Micro-tasks and crowd labour markets more broadly fundamentally shifted the international service economy [1] and had a disruptive effect on the digitisation of the workforce [2]. Still, at the centre of the debate on large-scale human computation projects is invariably a discussion of quality [3]. This is due in part to the fact that a suitable and scalable mechanism for the ex-ante detection of constantly underperforming contributors hasn't been presented, or, as provocatively posed by Roman in his note on crowdsourcing, there is no clear difference between "the wisdom of the crowd" and "the mob that rules" [4].

There are an ever increasing number of quality control measures but a gap exists in core theory to assist researchers and crowd market stakeholders, i.e., contributors and employers. Crowd labour markets and crowdsourcing exist in a state of 'perpetual beta,' defined by Kazman and Chen as an acceptance by requestors of 'on-going incompleteness" [5]. Tasks are structured so that the highest quality output is continually obtained and released [6]. Whilst requestors employ several mechanisms to assist in quality control, a holistic understanding of how well they work, why they work, and under which scenario they are working is missing. To date, commonly used tactics include qualification tests [7], pre-set qualifications [8]; trust models to determine the probability of diligent work [9]–[11]; hidden gold standard questions [12]; and the use of metrics such as solution acceptance [13] (see Related Work).

This has given rise to a suite of quality control measures leveraged in an almost á la carte fashion. The choice of one method over another obviously impacts the design, and execution of crowd work. Much of the literature concentrates on incremental improvements in quality, but does not establish a robust theory on the effects of quality measures to various platform stakeholders, i.e. contributor, employer, and platform provider. In this paper, we attempt to shed light on this perspective of quality control, such as to afford stakeholders making informed decisions surrounding the choice(s) of quality control measures.

One of the more challenging aspects from the perspective of organising platforms and the related policies is found in managing (constantly) underperforming contributors. This is in part due to a lack of insights into intent: Are contributors deliberately underperforming, or are they in need of additional coaching in order to perform to standard? This is obviously not an exhaustive list, and there may be a host of other answers to this question. Our study aims to illustrate that constantly underperforming contributors will not take on tasks that feature quality control mechanisms, irrespective of the quality control measure in place. Our objective is not to play one measure off against another, but rather contextualise their impact more holistically. Here we note that we do not use the term spammer, as we cannot predict the intention of our contributors.

Our research employs a 3 x 5 factorial experimental design of three task types with varying complexities

URI: https://hdl.handle.net/10125/59964
ISBN: 978-0-9981331-2-6

and five different quality control methods to measure the impact of quality control and task complexity on output quality (see Study Design). Our results indicate that the employed quality control method does not have a significant impact on response quality. In the experiment, it was sufficient to simply state that a qualification test is necessary to repel constantly underperforming contributors (see Results).

In our experiment most contributors were diligent, which has a strong implication on the management of crowd labour platforms. Constantly underperforming contributors by our definition (see Measurement) were only present in conditions with no quality control. This leads us to argue that expansive quality control support and applications are overused (see Conclusion). Simple, resource-light mechanisms are sufficient to assure quality results. In order to raise the global quality standards of crowdsourced contributions resources should be directed and dedicated to adequate training of contributors (see Implications).

2. Related Work

Quality control within crowdsourcing platforms appears in many ways, quality being one of the attributes of the crowd [18] [3]. Quality control is not only of interest to corporations and business: creative endeavours [14], policy and budget deliberations [15]–[17], open collaboration platforms [19], and the broader (scientific) community [4] stand to benefit as well. There are several other factors that may affect the quality of the work as per literature including the characteristics of the worker and demographics [20] or personality traits [21].

Literature in the field of crowdsourcing suggests various measures for assuring quality and authenticity. [22] tested the difference of user behaviour with honour codes compared to a serious warning message by conducting two experiments. Their first experiment subjects were college students taking an online test and the second experiment was Amazon MTurk. [22] found that replacing a traditional honour code with a strict warning decreases the chances of cheating in both student and MTurk environments by 50%. The authors state that informing the user regarding the negative consequences of an action by warning them results in lesser tendency of doing it. Kittur et al. [23] conducted two experiments to test MTurk as a user study platform. In their first experiment, they asked MTurkers to rate Wikipedia articles regarding their accuracy, writing, neutrality, structure, and the quality of the article. The workers were asked to fill a text box suggesting improvements to the article to verify if the user had really read the article. The authors did not find a correlation between the MTurkers ratings and the actual Wikipedia administrators. Their second experiment was the second version of the first one with slight modifications and additions in which they had both subjective and objective questions. Users were asked verifiable quantitative questions before rating the article. The users were asked to provide 4-6 keywords as a summary for the article. The results in experiment two demonstrated a significant correlation between the workers' ratings and the Wikipedia admin ratings. Kittur et al. [23] suggest that combining objective and subjective information gathering in user study tasks may be useful in micro-task markets.

Difallah et al. [24] discuss that crowdsourcing platforms do not share the worker's identity and they do not guarantee the quality of the work, which results in unreliability of the system. Cheaters were categorized a priori and posteriori and the authors discussed anti-adversarial techniques for encountering them. They suggest that sophisticated task formulation is a suitable obstacle for cheaters, however, it increases the burden on the requester whose main aim is to get the work done and suggest that applying traditional anti-spamming techniques such as CAPTCHA is a good option. They suggest that discouraging cheaters from doing a task is better than controlling the quality of completed tasks.

2.1 Pre-Selecting Contributors

Quality control mechanisms differ in their approach. In general, Kittur et al. [25] differentiate "up-front task design" and "post-hoc result analysis" as the two main methods to control work quality in a crowdsourcing context.

Crowdsourcing platforms provide the means for employers to pre-select contributors based upon specific task requirements or employer preferences. Geiger et al. [26] define pre-selection as "a means of ensuring a minimum ex-ante quality level of contributions." In other words, an employer will use a pre-selection process or test to mitigate the risk of poor quality solutions by screening potential contributors based upon the completion of some process that demonstrates certain knowledge or skills.

Oleson et al. [12] examine this process, which is typically performed via multiple-choice tests, and highlight as well as subsequently criticise a key assumption in this approach: that if the contributor passes the test, they will then perform the task well, even in the absence of direct or tangible incentives to do so. Similarly, if the contributor fails the test they may be banned from the task though not necessarily for the right reasons. This method is, however, simple to implement and also typically performs well. Pre-selection via qualification tests is also likely to act as a barrier for "scammer" contributors. However, diligent contributors may not select the task due to an increased

effort or missing credential on their part. Answers to a qualification test may also be shared amongst users, which reduces effectiveness.

2.2 Qualification Tests

Some platforms use a qualification test, to not only determine the abilities of a contributor, but also access and assess their basic properties, as this information is often not available to crowd employers. Stolee and Elbaum [27] and Chen et al. [28] are examples here. They state that a qualification can also capture demographic (and similar) properties of the contributor, for example geographical location. This does, however, massively distort the concept of a qualification if personal attributes are considered.

Similar to the basic notion of qualification tests are also initial screening questions based on reading attentiveness employed in order to minimize 'click-through' behaviours [29]. Such measures aim to ensure that contributors are dedicating significant attention to key elements of information, like the instructions.

2.3 In Task Quality Control

An alternative method proposed by Ipeirotis et al. [10] and Sheng et al. [30] is to infer a level of trust in the contributor via the accuracy of their solutions. Trust, however, quickly becomes a complex and nuanced topic highly specific to the context in which it is considered. Also as an inherently intangible and intransitive construct it is very difficult to measure quantitatively; key for approximating (automatically) a contributor's propensity for diligent or reliable work. Thus, Kern et al. [31] capture trustworthiness based on prior experience. They redundantly schedule tasks to multiple contributors to provide a basis to compare and estimate contributor reliability. This method demonstrated yielding high quality solutions. Yet without careful management the method is expensive in terms of redundantly issuing tasks (direct costs) and the additional effort needed to assess solution quality. Similarly, managing the crowd with respect to "rejected" answers can have other adverse effects, especially if the contributor has acted diligently.

Oleson et al. [12] propose the use of gold standard questions (frequently used on MTurk, for example) to assess solution quality and contributor ability. In their approach, subtasks with known solutions are injected into the task. The presence of these questions enables the accuracy of a given contributor to be estimated in task, and help improve the quality of their solutions by providing an explanation why the solution is incorrect. Contributors receive instant feedback on the accuracy of their performance. The approach, however, is limited to tasks that have a finite set of definite answers, and is inappropriate for tasks that rely on forms of subjectivity. However, such a mechanism provides a basis to also train a contributor, and enable self-evaluation of performance through feedback. The latter facilitates an integral element in the definition of competence: the evaluation of self-efficacy.

Quality Control, among others, is one of the dimensions in Quinn and Bedersen's classification dimensions of human computation [32]. The authors state that the users might cheat or sabotage the system even if they are motivated for participation. We believe that the rationale for subpar performance is the motivation being extrinsic rather than intrinsic. It is intrinsic motivation that plays a significant role as described in Self Determination Theory [33].

Ryan and Deci [33] define extrinsic motivation as "the performance of an activity in order to attain some separable outcome" the authors also discuss performing an activity to avoid punishment. Hence, we assume that the presence of any quality control procedure is efficient for quality as it invokes extrinsic motivation among contributors.

Reflecting on the different avenues of quality control, it is clear that much work has been undertaken in aligning the need for quality control and methods to underpin and support this need. Given the findings in recent literature and considering the idea of extrinsic motivation in Self-Determination Theory, we propose the following research question in order to evaluate quality assurance measures in crowdsourcing:

RQ: What is the relationship between quality control and perceived response quality in microtasks?

We explore if applying specific quality control methods have a significant impact on contributors' response quality, or simply whether just the announced the presence of a quality control method can prevent the contributors from underperforming.

3. Study Design

Our study had a three (task complexities) by five (quality control methods) factorial, between-group design. The experiment investigates three tasks of varying complexity. Following Allahbakhsh et al. [3], the effort for completing each task are as high or higher than for cheating, disincentivizing constant underperformance. We hypothesize the order of tasks in terms of complexity to be as follows *semantic* similarity (least complex), *question* answering (more complex), and text *translation* (most complex).

We repeated each task five times with different methods of quality control. For the first level of the *control* factor (*none*) we did not perform any quality control. For the second level (*fake*) we announced very prominently in the task description that we use introductory quizzes to check the qualification of contributors, yet contributors did not undertake a test. The third level (*intro*) announces an introductory quiz

and requires contributors to complete the quiz with 80% accuracy; akin to qualification tests. In the fourth level (*auto*) we added a basic machine learning (ML) system to estimate the quality of a response and report this estimate to contributors; akin to in task quality control measures. The system provides feedback on a three level scale (good, acceptable, unacceptable). Finally, in the fifth level (*wizard*) we replaced the ML-system by a human observer that decides the response quality. The scale was identical to the one used by the ML-system. Our objective with this measure is to represent an expert panel, reviewing each solution.

We recruited all contributors via *crowdflower,* as it allows international payments to be processed. We restricted our recruitment population to top-workers who were native English speakers to stimulate simple methods that can be used by any requester. To control possibly confounding variables, provide feedback, and perform our own quality control we redirected contributors to our own webpage. After completing the task contributors received a code that they use to receive their payment through the *crowdflower* interface. The user interface (Figure 1) was identical for all 15 (three by five) conditions. In all conditions, contributors were shown three examples of correctly solved tasks and a description of the task. We used the same interface to collect quality ratings from human judges.

We had a between-group design where each task had its own population. To ensure a this we used IP-tracking and browser fingerprinting to ensure that contributors do not contribute to more than one condition. There was no overlap among populations in the groups. To ensure contributor privacy only hashes of browser fingerprints and IP's were stored.

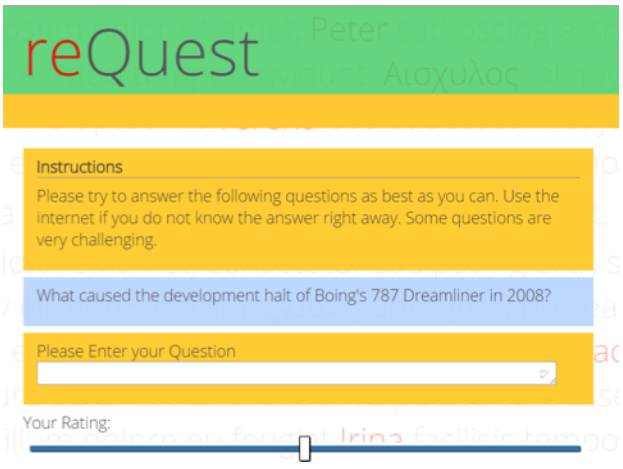

Figure 1: Crowdsourcing interface for the web-fragment annotation task. The interface is identical for all tasks. The rating slider (bottom) is only visible for our Raters when they judge the quality of a response.

3.1 Automated Feedback

The automated feedback system applied in the level *auto* of the *control* factor requires some explanation. Runge et al. [34] have shown that in some natural language tasks the quality of a response can be estimated with a high accuracy by a combination of the time needed to complete a single request and the numbers of characters typed. Although the values of both variables and their meaning differ from task to task, a ML classifier is able to learn the relationship between the two variables (features) and the response quality with minimal training data.

For our *auto* level, we classify responses into three different classes (good, acceptable, unacceptable) using a random forest classifier [35]. Supervised classifiers need labelled training data. We classified 90 responses of each task by hand. We randomly selected responses and classified them into the three classes until there were 30 samples per class. We stratified the training data randomly, selecting exactly 30 samples per class.

For the experiment a random forest classifier was chosen, as tree-based classifiers are less sensitive to outliers and unbalanced sample sets [36]. In the given tasks, it is likely that we encounter outliers such as a contributor opening a task and leaving their working place for a while. Classifiers such as support vector machines are more sensitive to such outliers. Our classifier generated 10 random trees using Gini impurity [37] as the split criterion, built using the python sklearn package [38].

When the classifier estimates the response quality to be unacceptable we show a general warning that the response might need revision. If the response was acceptable, we did not show a message. For good responses, a message stating that the response was of good quality is shown. Messages appeared as a red text immediately after a contributor responded to a request.

3.2 Measurements

We consider two independent variables: the quality *control* method and *task complexity* as well as one dependent variable: perceived response *quality*. To measure perceived response quality, we asked two human judges to rate each response on a scale from 0.0 (low quality) to 1.0 (high quality) in 10 increments. We calculated the average perceived response quality for each contributor as our measurement for *quality*. We consider contributors with an average perceived response quality below 0.6 as constantly underperforming, i.e. 40% unacceptable responses.

Judges saw the initial request and answer. Additionally, judges had a slider to rate the response quality (see Figure 1). The interface did not show the rating of our automated feedback system. We ensured

that the process was blind. We randomly selected responses from all conditions and judges did not know the condition of a response. These judges were not involved in generating the training data for the automated feedback nor did they participate in the *wizard* conditions. We recruited the judges' offline.

We measure and report the agreement between judges using Krippendorff's Alpha [39]. Additionally, we measure the correlation between our ML-systems prediction and our human judges. As our data violates the assumptions of the Pearson Product-Moment correlation we use Spearman's ρ.

Furthermore, the three tasks are tested for instruction clearness and contributor satisfaction using the build in metrics provided by *crowdflower*. Upon completion of a task, contributors can take a satisfaction survey. Contributors score the task on a 0-5 scale for overall satisfaction, instruction clearness, test question fairness, payment, and ease of job. Results of these quizzes are reported with each task.

4. Procedure

We collected all data for three independent tasks from the domain of natural language processing. The main interface for contributors is identical for all tasks. Figure 1 shows a screenshot of the user interface for the question-answering task. Table 1 shows the distribution of our contributors by level of quality control method and task complexity.

	None	Fake	Intro	Auto	Wizard
Semantic	17	19	17	18	19
Question	19	17	16	19	18
Translation	16	17	18	19	20

Table 1: Distribution of contributors over all 15 conditions.

4.1 Word-based Semantic Similarity

Semantic similarity plays an important role for many natural language processing tasks, especially word sense disambiguation and information retrieval [40], [41]. Humans are better than algorithms at rating semantic similarity between two words [7]. Involving paid online contributors can reduce costs, but the response quality is harder to predict. Constantly under-performing contributors are still an issue for such tasks [11]. Different algorithmic approaches do exist [42]–[44] but are not yet able to reproduce human level results [45]. The task issued in this treatment is itself not very complex, only requiring a good command of English. To ensure this, we restricted contributor's origin to be in the US, UK, or Canada. We further restricted the task using a standard dataset [46] consisting of 353 word pairs. In the experiment, we recruited 90 contributors and collected ~9,500 responses on the 353 word pairs.

4.2 Question Answering

Understanding natural language is still a challenging field for artificial systems [47]. Answering questions given in natural language or finding relevant search results to these questions are, despite the recent success of systems such as IBM Watson [48], unsolved challenges [49], [50]. As standard datasets for question answering seem too easy for human annotators with access to the internet, we designed a set of 50 questions so that using the question as a search string will not reveal the correct answer right away.

We randomly selected 10 questions to be test questions for conditions with an introductory test (*Intro, Auto, Wizard*). We designed sets of possible answers to these 10 test questions by hand. Each answer set had ~10 answers from at least three different people. Answers were collected off-line from students and members of our research group. The response quality of a contributor is estimated by the semantic similarity between the contributor's response and our exemplary answers. We take the highest similarity value as an estimate of quality. The method is calibrated by testing each of the handmade answers against the remaining answers in each set. The average similarity of answers on a scale from 0.0 (no similarity) to 1.0 (perfect similarity) is 0.65 (SD: 0.25). Responses within a margin of one standard deviation were considered acceptable.

Each contributor could answer up to 80 questions. We collected 5,089 responses (57 on average) from 89 contributors on *crowdflower*. We collected 1,017 responses on average for each *control* level.

4.3 Text Translation

Text translation is a demanding task even for humans as in-depth knowledge of two different domains, the target and the source language, is required. Various approaches exist; applying crowdsourcing to translation targeted paraphrasing [51] and iterative collaboration between monolingual users [52] are two examples. Other common approaches utilize mono- or bilingual speakers to proofread and correct Machine Translation results [53]. For our experiment, we use a popular Wikipedia article in German on the Brandenburg Gate. Native speakers of German prepared a set of sentences from this article. For the set, we took the first 150 sentences from the respective article. Headlines, incomplete sentences and sentences that contained words in a strong dialect were removed. We requested translations for the remaining sentences from contributors via crowdflower. As the target language was English we used the same quality prediction method for conditions that included a pre-

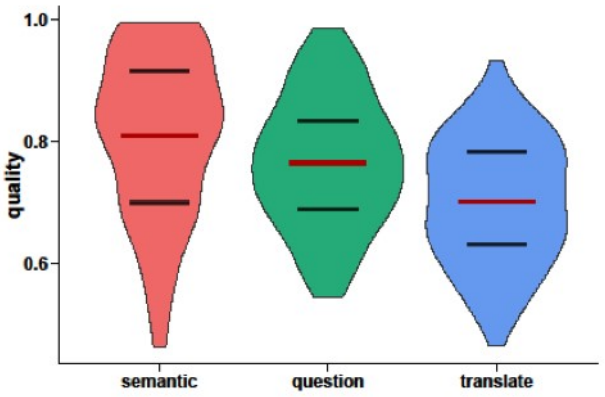

Figure 2: Task complexity affects response quality. The most complex task text translation (right) has a significantly lower average response quality than the more simplistic semantic similarity task (left) and the question answering task (middle). The figure shows a violin plot combining a boxplot and a kernel density plot. Thick dark lines indicate 1st and 3rd quartiles the red lines population means.

test as for the question answering task. Each contributor could translate up to 100 sentences. We collected 2,119 translations for the Vietnamese set and 2,002 translations for the German set (total 4121) from 90 contributors (46 on average). We collected 825 sentences on average in each *control* condition.

5. Results

Before we analyse our data, we want to ensure that our presumption that the three different tasks have a distinct complexity is reasonable. We indeed found that the response quality is significantly lower for complex tasks. This indicates that the tasks do differ in their complexity. This is in line with the self-assessment of contributors through *crowdflowers* satisfaction survey. We found that *Ease Of Job* negatively correlates with our presumed complexity ranking. The correlation is significant with $p < 0.001$. Table 2 shows the results of the satisfaction survey.

	Satisfaction	*Clarity*	*fairness*	*Payment*	*Ease*
Similarity	3.8	3.8	3.7	4.5	4.3
Question	3.6	3.4	3.5	4.1	3.7
Translate	3.7	3.9	3.3	4.4	3.1

Table 2: Results of the self-assessment. From left to right the columns refer to overall satisfaction, instruction clearness, test question fairness, payment, and ease of job. It is not possible to calculate a SD as crowdflower only offers aggregated data.

Then we ensure that our metric is reasonable. Perceived quality is used as this measure allows investigating quality over different tasks. Table 3 shows that our judges have a substantial agreement on quality throughout all tasks.

Before testing our results for significance, we ensured that our data is suitable for parametric tests. We used the Shapiro-Wilk test for normality [54] for each condition and did not find significant differences from a normal distribution.

	Participants	*Judges*	*Krippendorff's α*
Similarity	90	2	0.808
Question	89	2	0.838
Translate	90	2	0.815

Table 3: Inter-rater agreement on perceived response quality. The results are homogenous for all three tasks and indicate a substantial agreement between our judges.

As we have different numbers of contributors in our conditions, we also verified that our conditions have equal variance for the dependent variable prior to executing an analysis of variance (ANOVA). As the distributions do not differ significantly from normal distributions we use Bartlett's test for homoscedasticity (equal variance) [55]. We found that the variance does not differ significantly between our conditions $t(4) = 2.764$, $p = 0.598$. As our data does not hold evidence that it violates the assumptions of the ANOVA, we analyse main and interaction effects with a two-way ANOVA to compare the effect of quality *control* and *task* complexity on the independent variable perceived response *quality*. Table 4 shows these results.

	df	*SS*	*MS*	*F*	*p*	*sig.*
(C)ontrol	4	1.036	0.259	28.988	0.001	***
(T)ask	2	0.557	0.279	31.165	0.001	***
CxT	8	0.220	0.028	3.082	0.002	**
Residuals	254	2.270	0.009			

Table 4: ANOVA results of main and interaction effects. The first row shows the effect of the quality control method. The second row the effect of the task. The third row shoes the interaction effect between both factors.

From the ANOVA results, we conclude that *task* complexity as well as the used quality *control* method have a significant influence on the perceived response *quality*. Furthermore, we found a significant interaction between both factors. We use Welch Two Sample t-test with Holm-Bonferroni correction as our post hoc comparison method. Table 5 presents differences in levels of the *control* factor.

Task Complexity Affects Response Quality

We analyse effects for each level of the *task complexity* factor, assuming that the average response quality deteriorates with higher complexity tasks. As seen in Table 6 and Figure 2 this assumption holds. Although this may seem obvious it substantiates the initial

assumption on task complexity. The Pearson moment correlation is 1.0 with an associated $p < 0.001$.

Comp.	*M1*	*SD1*	*M2*	*SD2*	*T*	*df*	*p*	*Sig.*
none fake	0.63	0.09	0.80	0.11	-8.21	100	0.00	***
none intro	...	...	0.79	0.12	-7.72	97	0.00	***
none auto	...	...	0.78	0.13	-7.67	105	0.00	***
none wiz.	...	...	0.79	0.13	-8.17	106	0.00	***
fake intro	0.80	0.11	0.79	0.12	0.44	102	0.66	
fake auto	...	...	0.78	0.13	0.74	106	0.46	
fake wiz.	...	...	0.79	0.13	0.25	107	0.80	
intro auto	0.79	0.11	0.78	0.11	0.29	104	0.77	
intro wiz.	...	...	0.79	0.13	-0.20	105	0.85	
auto wiz.	0.78	0.11	...	...	-0.50	111	0.62	

Table 5: Welch two sample t-tests with Holm correction comparing all levels of the quality control factor.

Comp.	*M1*	*SD1*	*M2*	*SD2*	*T*	*df*	*p*	*Sig.*
Sem. Quest.	0.81	0.13	0.77	0.11	2.45	169	0.02	*
Sem. Trans.	...	...	0.70	0.10	6.07	167	0.00	***
Quest Trans.	0.77	0.11	0.70	0.10	4.10	177	0.00	***

Table 6: Results of Welch two sample t-tests with Holm correction. Line 1 compares level semantic to level question of the task complexity factor. Line 2 compares level semantic translation and the line three question to translation

The results indicate that there is a significant difference between the levels *none* of *control* and the other four levels. The resulting p-values are below the 0.001 alpha-level as seen in Table 5. Other levels do not differ significantly. Table 7 shows means and standard deviations between all levels of our two factors. Figure 3 further illustrates that the finding is constant for all tested tasks.

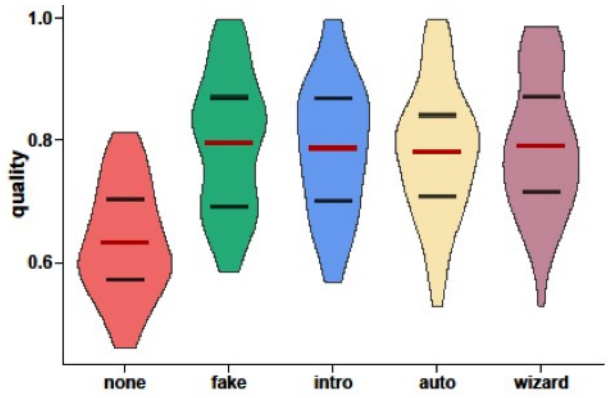

Figure 3: Quality control affects response quality only if there is no quality control at all. The differences in means between quality control methods are not significant.

	Semantic		Question		Translation	
	M	SD	M	SD	M	SD
none	0.62	0.09	0.68	0.09	0.60	0.08
intro	0.84	0.11	0.78	0.09	0.74	0.11
fake	0.85	0.10	0.81	0.11	0.72	0.09
auto	0.89	0.07	0.76	0.06	0.70	0.10
wizard	0.83	0.11	0.81	0.13	0.73	0.07

Table 7: Means and standard deviations for perceived quality. Rows contain the five different quality control methods and columns the different tasks.

We also investigated the proportion of constantly underperforming contributors (a contributor below a quality level of 0.6). We found that in all no-quality control conditions we had a substantial amount of contributors (N = 22) with an average response quality below 0.6. In all other conditions combined, we found 11 contributors under this threshold. The proportion of underperforming contributors in the *none* conditions is 0.42. Compared to the other conditions with a proportion of only 0.05 this is value is extremely high.

In the *auto* level of the quality control factor a ML-System predicted the response quality of contributors based on two features (number of characters typed and time needed to complete a request). To estimate the quality of this prediction we calculated the correlation between our ML-systems prediction and the average perceived quality. The ML-system rated responses on a scale with three ordered values (unacceptable (1); acceptable (2); good (3)). As this scale is ordinal and violates the assumptions of Pearson's Product-Moment correlation we analysed the correlation using Spearman's ρ. We found a substantial correlation between the predictions and the average perceived quality of our human judges $\rho\ (937020) = 0.71$, $p < 0.001$. The correlation between the two human judges in comparison is $\rho\ (463061) = 0.85$, $p<0.001$. In contrast, the human raters who replaced the ML-system in our wizard condition achieved a correlation of $\rho\ (705574) = 0.78$, $p<0.001$.

6. Conclusion

In this paper, we investigated the effect of different quality control methods on the response quality of contributors for tasks of varying complexity. We established their differing complexity and confirmed the order to be as follows semantic similarity (least complex), question answering (more complex), text translation (most complex).

We found that constantly underperforming contributors (by our definition contributors with less than 40% acceptable responses) are almost not present in all conditions of our experiment when a quality control method is in place. We however found a substantial amount of constantly underperforming contributors (almost 45%) in our control conditions (*none*) without a quality control method.

Only mentioning a required introductory test (without actually doing the test, the *fake* level of the *control* factor) was sufficient to achieve the same response quality as the quality control methods. Even immediate human generated feedback was not able to raise response quality above the level of this faked introductory test. As hypothesized, the response quality does not differ across the different quality control methods. It only differs significantly between the *none* conditions (M = 0.63, SD = 0.03) and conditions with quality control (M = 0.79, SD = 0.05). This is an increase of more than 25% in response quality.

We can therefore conclude that constantly underperforming contributors are aware of the fact that their contribution might fall short of required quality standards when taking a task. This also implies that very basic quality control methods are sufficient to promote diligent work. Yet, it is debatable if our fake introductory test would keep these results over time. It is very likely that contributors realize that the tests are not conducted, and it is also known that contributors share task information amongst themselves. However, we also demonstrated that extremely simple ML methods with task independent features as proposed by Krause et al. [50] can predict response quality on the fly. Such methods may provide quality control for tasks similar to the ones explored in this paper.

7. Implications

The core contributions of this study address platforms and requestors. We argued that multiple quality control measures exist, but effectiveness of said mechanisms at a meta-level is still under-addressed. This work addresses that gap, extending existing knowledge on the comparative effectiveness of various quality control regimes. Our findings suggest that increasingly complex, resource-intensive quality assurance mechanisms do not have better performance than simple mechanisms. As shown in this work, after an even basic controlling for response quality, underperformance per task drops considerably. Investments in simple mechanisms should be prioritised above resource-intensive mechanisms.

Our work aspires to address the status of contributors as well. Diligent but underperforming contributors can exist for many reasons, and are likely to be wrongly classified as spammers. At the same time, response quality degrades with increasing task complexity. This points to the need for suitable training and developmental materials. Our argument is simple: rather than investing in post-hoc quality control, investing ad-hoc in training and skill development should increase quality globally. A natural extension is the creation and validation of credential regimes, something missing and drastically needed for underpinning and securing contributors' rights in crowd labour markets [3], [13].

Our findings indicate that as discussed in Self-Determination Theory, the presence of any quality control method activates the extrinsic motivation – avoiding punishment - in contributors. Returning to the question of measure mechanics vs. the perception that workers have towards the broader notion of a quality measure being in place, based upon our initial results the latter maybe more impactful. This should not be interpreted as a sweeping dismissal of research into quality measures and their mechanics. Instead, it should highlight that the fundamental components of competence, where the evaluation of self-efficacy is just one, aligned with the reputation systems that crowd platforms employ act as a significant deterrent for underperforming workers. In other words, participation in such a task is expressing the belief of providing a valid solution by the contributor. Further study is, however, needed to more thoroughly evaluate this and enable a more rigorous theory on the interplay between quality control measures and the associated policies in crowd labour markets.

8. Limitations and Future Work

While a 3x5 factorial model is sizable, future work should cover more quality control mechanisms to assure the transferability of these results. Furthermore, it has yet to be seen if tasks in other domains than natural language processing yield similar results.

We recognize that our minimal control mechanism (*fake*) without enforcement is not sustainable - contributors can and will realize that no quality control has in fact been enforced. A sustainable and low cost mechanism to elevate the performance of diligent but underperforming contributors must be developed and tested to complete the scope of this research.

A worthy area of future research is support systems for those who worked diligently but are still underperforming. This is both for the requestor's side (i.e., clear task description writing) and contributor's side (i.e., developmental educational materials) [56].

Particularly worthwhile would be the investigation of monetary incentivization of contributors' education (see e.g., [57], [58]). Monetized education-based tasks

could create the scenario that contributors are both learning to complete more and more complex tasks, while gaining skills and funding to be applied in their offline lives. An envisioned mechanism for this could be Massively Open Online Courses, where contributors register for the course to learn increasingly complex skills, and are financially rewarded with successful task mastery. Realized in its full depth and scope, this progressive step would comprehensively enhance of both crowdwork from a quality perspective and the overall, real life skillset of the contributors.

9. References

[1] O. Kässi and V. Lehdonvirta, "Building the Online Labour Index : A Tool for Policy and Research," in *CSCW 2016 workshop on The Future of Platforms as Sites of Work, Collaboration and Trust*, 2016.

[2] V. del Rosal, *Disruption: Emerging Technologies and the Future of Work*. CreateSpace Independent Publishing, 2015.

[3] M. Allahbakhsh, B. Benatallah, and A. Ignjatovic, "Quality Control in Crowdsourcing Systems," *IEEE Internet Comput.*, vol. 17, no. 2, pp. 76–81, 2013.

[4] D. Roman, "Crowdsourcing and the Question of Expertise," *Commun. ACM*, vol. 5, no. 1, p. 1610258, 2009.

[5] R. Kazman and H. Chen, "The Metropolis Model: A New Logic for Development of Crowdsourced Systems," *Commun. ACM*, vol. 52, no. 7, pp. 76–84, 2009.

[6] T. I. M. Straub, H. Gimpel, and F. Teschner, "The Negative Effect of Feedback on Performance in Crowd Labor Tournaments," pp. 2012–2015, 2014.

[7] N. Batram, M. Krause, and P. Dehaye, "Comparing Human and Algorithm Performance on Estimating Word-based Semantic Similarity," in *SoHuman'14 Proceedings of the 3rd International Workshop on Social Media for Crowdsourcing and Human Computation*, 2014, pp. 131–139.

[8] C. Sarasua and M. Thimm, "Microtask available, send us your CV!," in *Proceedings - 2013 IEEE 3rd International Conference on Cloud and Green Computing, CGC 2013 and 2013 IEEE 3rd International Conference on Social Computing and Its Applications, SCA 2013*, 2013, pp. 521–524.

[9] J. Wang, P. Ipeirotis, and F. Provost, "Quality-Based Pricing for Crowdsourced Workers," pp. 1–46, 2013.

[10] P. G. Ipeirotis, F. Provost, and J. Wang, "Quality management on amazon mechanical turk," in *HComp'10 Proceedings of the ACM SIGKDD Workshop on Human Computation*, 2010, pp. 0–3.

[11] M. Krause and R. Porzel, "It is about time," in *CHI '13 Extended Abstracts on Human Factors in Computing Systems on - CHI EA '13*, 2013, p. 163.

[12] D. Oleson, A. Sorokin, G. Laughlin, V. Hester, J. Le, and L. Biewald, "Programmatic Gold: Targeted and Scalable Quality Assurance in Crowdsourcing," *HComp'11 Proc. AAAI Work. Hum. Comput.*, pp. 43–48, 2011.

[13] C. Dukat and S. Caton, "Towards the competence of crowdsourcees: Literature-based considerations on the problem of assessing crowdsourcees' qualities," in *Proceedings - 2013 IEEE 3rd International Conference on Cloud and Green Computing, CGC 2013 and 2013 IEEE 3rd International Conference on Social Computing and Its Applications, SCA 2013*, 2013, pp. 536–540.

[14] R. M. Araujo, "99designs: An Analysis of Creative Competition in Crowdsourced Design," in *First AAAI conference on Human computation and crowdsourcing*, 2013, pp. 17–24.

[15] G. Smith, R. C. R. Jr, and J. Gastil, "The Potential of Participedia as a Crowdsourcing Tool for Comparative Analysis of Democratic Innovations," *Policy & Internet*, vol. 7, no. 2, 2015.

[16] J. Prpic, A. Taeihagh, and J. Melton, "The Fundamentals of Policy Crowdsourcing," *Policy & Internet*, vol. 7, no. 3, pp. 340–361, 2015.

[17] C. Niemeyer, T. Teubner, M. Hall, and C. Weinhardt, "The Impact of Dynamic Feedback and Personal Budgets on Arousal and Funding Behaviour in Participatory Budgeting," *Gr. Decis. Negot.*, 2018.

[18] J. Prpić and P. Shukla, "Crowd science: Measurements, models, and methods," *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, vol. 2016–March, pp. 4365–4374, 2016.

[19] D. Friess and C. Eilders, "A Systematic Review of Online Deliberation Research," *Policy & Internet*, vol. 7, no. 3, pp. 319–339, 2015.

[20] G. Kazai, "The Face of Quality in Crowdsourcing Relevance Labels : Demographics , Personality and Labeling Accuracy," pp. 3–6, 2012.

[21] G. Kazai, J. Kamps, and N. Milic-Frayling, "Worker types and personality traits in crowdsourcing relevance labels," *Proc. 20th ACM Int. Conf. Inf. Knowl. Manag. - CIKM '11*, p. 1941, 2011.

[22] H. Corrigan-Gibbs, N. Gupta, C. Northcutt, E. Cutrell, and W. Thies, "Deterring Cheating in Online Environments," *ACM Trans. Comput. Interact.*, vol. 22, no. 6, pp. 1–23, 2015.

[23] A. Kittur, E. H. Chi, and B. Suh, "Crowdsourcing User Studies With Mechanical Turk," 2008.

[24] D. E. Difallah, G. Demartini, and P. Cudr??-Mauroux, "Mechanical cheat: Spamming schemes and adversarial techniques on crowdsourcing platforms," *CEUR Workshop Proc.*, vol. 842, pp. 20–25, 2012.

[25] A. Kittur *et al.*, "The Future of Crowd Work," in *Proceedings of the 2013 conference on Computer supported cooperative work - CSCW '13*, 2013, p. 1301.

[26] D. Geiger, S. Seedorf, R. Nickerson, and M. Schader, "Managing the Crowd : Towards a Taxonomy of Crowdsourcing Processes," in *Proceedings of the 17th Americas Conference on Information Systems*, 2011, pp. 1–11.

[27] K. T. Stolee and S. Elbaum, "Exploring the use of crowdsourcing to support empirical studies in

software engineering," in *Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, 2010, p. 35.

[28] J. J. Chen, N. J. Menezes, A. D. Bradley, and T. A. North, "Opportunities for Crowdsourcing Research on Amazon Mechanical Turk," *Interfaces (Providence).*, vol. 5, no. 3, 2011.

[29] A. J. Berinsky, G. Huber, and G. S. Lenz, "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk," *Polit. Anal.*, vol. 20, no. 3, pp. 351–368, Mar. 2012.

[30] V. S. Sheng, F. Provost, and P. G. Ipeirotis, "Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers," in *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, p. 614.

[31] R. Kern, H. Thies, and G. Satzger, "Statistical Quality Control for Human-based Electronic Services," in *Proceedings of Service-oriented computing: ICSOC 2010*, 2010, pp. 1–17.

[32] A. J. Quinn and B. B. Bederson, "Human Computation: A Survey and Taxonomy of a Growing Field," *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, pp. 1403–1412, 2011.

[33] R. Ryan and E. Deci, "Self-determination theory and the facilitation of intrinsic motivation," *Am. Psychol.*, vol. 55, no. 1, pp. 68–78, 2000.

[34] N. Runge, N. Kilian, J. Smeddinck, and M. Krause, "Predicting Crowd-Based Translation Quality with Language-Independent Feature Vectors," *Work. Twenty- …*, pp. 114–115, 2012.

[35] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[36] D. Cieslak and N. Chawla, "Learning decision trees for unbalanced data," in *European Conference on Machine Learning*, 2008.

[37] L. Breiman, "Technical note: Some properties of splitting criteria," *Mach. Learn.*, vol. 24, no. 1, pp. 41–47, Jul. 1996.

[38] F. Pedregosa and G. Varoquaux, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[39] K. Krippendorff, "Estimating the Reliability, Systematic Error and Random Error of Interval Data," *Educ. Psychol. Meas.*, vol. 30, no. 61, pp. 61–70, 1970.

[40] J. Feng, Y. Zhou, and T. Martin, "Sentence Similarity based on Relevance," in *Proceedings of the 12th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU '08)*, 2008, pp. 832–839.

[41] R. Navigli, "Word sense disambiguation: A survey," *ACM Comput. Surv.*, vol. 41, p. 10, 2009.

[42] M. Strube and S. P. Ponzetto, "WikiRelate ! Computing Semantic Relatedness Using Wikipedia," in *AAAI*, 2006, pp. 1419–1424.

[43] D. Yang and D. M. W. Powers, "Measuring Semantic Similarity in the Taxonomy of WordNet," *Reproduction*, pp. 315–322, 2005.

[44] P. Resnik, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, 1995, p. 6.

[45] K. Radinsky, E. Agichtein, E. Gabrilovich, and S. Markovitch, "A word at a time: computing word relatedness using temporal semantic analysis," in *Proceedings of the 20th International World Wide Web Conference WWW'11*, 2011, pp. 337–346.

[46] L. Finkelstein *et al.*, "Placing Search in Context : The Concept Revisited," pp. 406–414, 2001.

[47] M. Krause, "Designing Systems with Homo Ludens in the Loop," in *Handbook of Human Computation*, 1st ed., P. Michelucci and K. Greene, Eds. New York, NY, USA: Springer New York, 2014, pp. 393–409.

[48] D. Ferrucci, E. Brown, J. Chu-Carroll, and J. Fan, "Building Watson: An overview of the DeepQA project," *AI Mag.*, vol. 31, no. 3, pp. 59–79, 2010.

[49] H. Aras, M. Krause, A. Haller, and R. Malaka, "Webpardy: Harvesting QA by HC," in *HComp'10 Proceedings of the ACM SIGKDD Workshop on Human Computation*, 2010, pp. 49–53.

[50] M. Krause, *Homo Ludens in the Loop: Playful Human Computation Systems*. Hamburg, Germany: tredition GmbH, Hamburg, 2014.

[51] P. Resnik, H. Chang, O. Buzek, and B. B. Bederson, "Using Monolingual Human Computation to Improve Language Translation via Targeted Paraphrase," in *HComp'10 Proceedings of the ACM SIGKDD Workshop on Human Computation*, 2010.

[52] H. Chang, B. B. Bederson, and P. Resnik, "MonoTrans2: An Asynchronous Human Computation System to Support Monolingual Translation," *hcil.cs.umd.edu*. pp. 2–5, 2010.

[53] O. F. Zaidan and C. Callison-burch, "Crowdsourcing translation: professional quality from non-professionals," in *Proceedings of ACL 2011*, 2011, pp. 1220–1229.

[54] J. P. Royston, "An Extension of Shapiro and Wilk's W Test for Normality to Large Samples," *J. R. Stat. Soc. Ser. C-Applied Stat.*, vol. 31, pp. 115–124, 1982.

[55] M. Bartlett, "Properties of sufficiency and statistical tests," *Proc. R. Stat. Soc. Ser.*, vol. A, no. 160, pp. 268–282, 1937.

[56] M. Streuer, M. Krause, M. Hall, and S. Dow, "On-the-Job Learning for Micro-Task Workers," in *HCOMP*, 2017.

[57] M. Krause, M. Hall, J. J. Williams, P. Paritosh, J. Prpic, and S. Caton, "Connecting Online Work and Online Education at Scale," in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2016, pp. 3536–3541.

[58] R. Suzuki, N. Salehi, M. S. Lam, J. C. Marroquin, and M. S. Bernstein, "Atelier: Repurposing Expert Crowdsourcing Tasks as Micro-internships," in *Chi 2016*, 2016.